BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 4: e59412
doi: 10.3897/biss.4.59412

OPEN ACCESS

Conference Abstract

Outlier Detection at GBIF Using DBSCAN

John Thomas Waller ‡

‡ GBIF, Copenhagen, Denmark

Corresponding author: John Thomas Waller (jwaller@gbif.org)

Received: 07 Oct 2020 | Published: 08 Oct 2020

Citation: Waller JT (2020) Outlier Detection at GBIF Using DBSCAN. Biodiversity Information Science and Standards 4: e59412. https://doi.org/10.3897/biss.4.59412

Abstract

Geographic outliers at GBIF (Global Biodiversity Information Facility) are a known problem. Outliers can be errors, coordinates with high uncertainty, or simply occurrences from an undersampled region. Often in data cleaning pipelines, outliers are removed (even if they are legitimate points) because the researcher does not have time to verify each record one-by-one. Outlier points are usually occurrences that need attention. Currently, there is no outlier detection implemented at GBIF and it is up to the user to flag outliers themselves.

DBSCAN (a density-based algorithm for discovering clusters in large spatial databases with noise) is a simple and popular clustering algorithm. It uses two parameters, (1) distance and (2) a minimum number of points per cluster, to decide if something is an outlier. Since occurrence data can be very patchy, non-clustering distance-based methods will fail often Fig. 1. DBSCAN does not need to know the expected number of clusters in advance. DBSCAN does well using only distance and does not require some additional environmental variables like Bioclim.

© Waller J. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

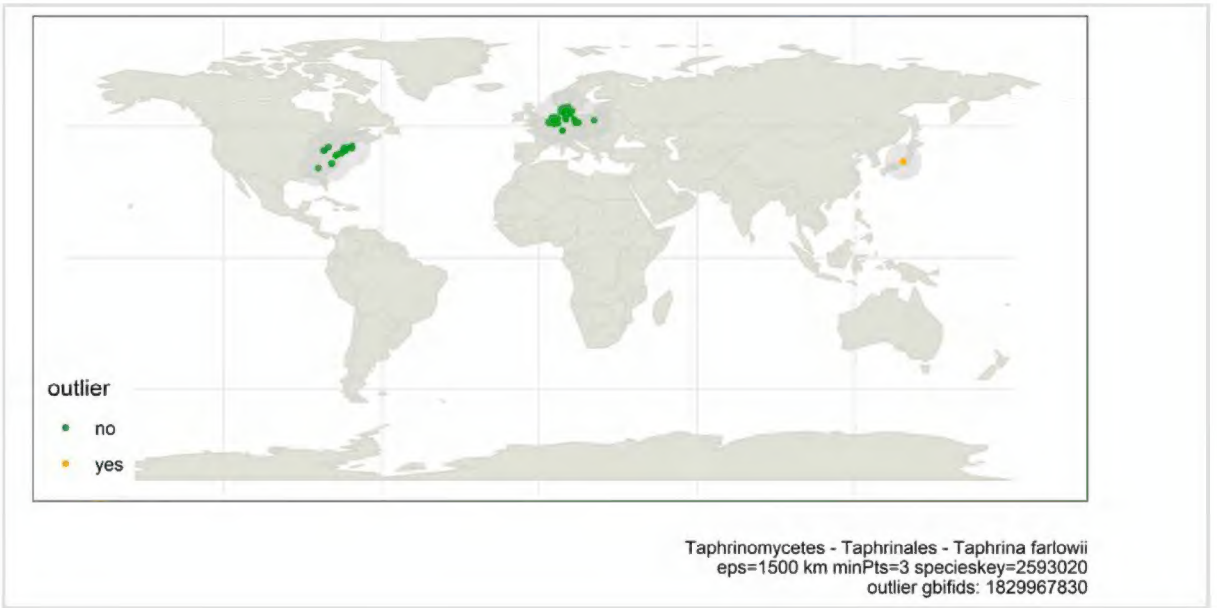

Figure 1.

This example shows that DBSCAN is able to cluster effectively while flagging points with low additional support in Japan (outlier point).

Advanatages of DBSCAN :

- Simple
- Easy to understand
- Only two parameters to set
- Scales well
- No additional data sources needed
- Users would understand how their data was changed

Drawbacks :

- Only uses distance
- Must choose parameter settings
- Sensitive to sparse global sampling
- Does not include any other relevant environmental information
- Can only flag outliers outside of a point blob

Outlier detection and error detection are different. If your goal is to produce a system with no false positives, it will fail. While more complex environmentally-informed outlier detection methods (like reverse jackknifing (Chapman 2005)) might perform better for certain examples or even in genreal, DBSCAN performs adequately on almost everything despite being very simple.

Currently I am using DBSCAN to find errors and assess dataset quality. It is a Spark job written in Scala (github). It does not run on species with lots of (>30K) unique latitude-longitude points, since the current implementation relies on an in-memory distance matrix. However, around 99% of species (plants, animals, fungi) on GBIF have fewer than >30K unique lat-long points (2,283 species keys / 222,993 species keys). There are other implementations (example) that might scale to many more points.

There are no immediate plans to include DBSCAN outliers as a data quality flag on GBIF, but it could be done somewhat easily, since this type of method does not rely on any external environmental data sources and already runs on the GBIF cluster.

Keywords

data quality, georeference

Presenting author

John Thomas Waller

Presented at

TDWG 2020

References

- Chapman (2005) Principles and Methods of Data Cleaning: Primary Species and Species-Occurrence Data. 1. GBIF URL: https://www.gbif.org/document/80528/principles-and-methods-of-data-cleaning-primary-species-and-species-occurrence-data